A Data-driven Approach to Optimize Bounds on the Capacity of the Molecular Channel

Francesca Ratti, Gabriele Scalia, Barbara Pernici and Maurizio Magarini
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, I-20133 Milano, Italy
Email: {francesca.ratti, gabriele.scalia, barbara.pernici, maurizio.magarini}@polimi.it

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322078

***Abstract*—The study of channel capacity is a well-known problem in Digital Communication (DC) systems. Most of the channel models used to evaluate capacity consider the additive white Gaussian noise as the sole impairment. Analytical formulas for lower and upper bounds have been obtained considering such a statistical characterization and different constraints for the transmitted signal. The field of Molecular Communication (MC) shows several analogies with DC systems. However, to the best of our knowledge, it is not possible to determine a statistical model characterizing an MC channel that considers the nonlinear effects present in the system. This paper aims to develop a *data-driven* methodology that, starting from *in-silico* or *in-vitro* experiments, allows estimating bounds on the constrained channel capacity of any biological system and the corresponding distribution of the source message, e.g., finite concentration levels of a protein. As experiments are time consuming, the method includes a machine learning-based data augmentation step. Our proposal is illustrated for a biological circuit composed of two prokaryotic cells. Results highlight fast and stable convergence of the algorithm to tight capacity bounds.**

I. Introduction

The field of Molecular Communication (MC) has gained much interest in the recent years as an emerging discipline directly inspired by natural communications between cells in biology [1], [2]. Understanding living cells from an information and communication theoretical perspective is one of the challenges to gain insights into the fundamentals of MC system engineering [3]. A mathematical modeling exists for many of the phenomena occurring in an MC system. For example, the number of absorbed molecules by a cell in a time-invariant MC channel can be modelled with a Binomial distribution [4], [5], or even with Gaussian or Poisson one.

An open research problem in MC is that of finding optimal ways to transfer information on constrained channels considering the effects of processes, e.g., biochemical reactions, that cannot be characterized by an analytical formulation. Several works address the problem of optimizing information transfer on the MC channel. For example, in [6] the authors analyzed the channel capacity considering *both* diffusion-based channel and ligand-based receiver, while in [7] the authors provided a closed-form expression for the information capacity of an MC system with a noisy channel. In [8] a different approach is taken, where the authors used enzymatic reaction cycles to improve the upper bound on the mutual information for a diffusion-based MC system. The work in [9] presented an analysis of the channel capacity in diffusive MC by considering inter-symbol interference from all the previous time slots and the channel transmission probability in each time slot. A common aspect of these works is that of relying on an analytical model to describe the channel between the transmitter and the receiver.

We pursue a similar goal, but we develop a methodology for systems for which an analytical formulation is not available. As in [10], the objective of this work is to minimize the upper bound and to maximize the lower bound on the channel capacity with a constraint on the amplitude of the input message, starting from the calculation of the Mutual Information (MI), which is usually adopted to measure the information exchange between source and destination. We propose a *data-driven* approach based on an iterative method that maximizes the lower bound on the channel capacity defined as a difference of entropies and that minimizes the upper bound derived by the definition of Kullback-Leibler divergence. The proposed methodology is applicable to all the communication systems for which a statistical characterization of the channel does not exist, but it is instead possible to generate a dataset of source messages and corresponding received signals. This happens in a variety of fields including, for example, psychology, signal processing or finance [11]. In the molecular biology field, it is possible to generate such a dataset from *in-vitro* or *in-silico* experiments. As in-silico/in-vitro experiments are time consuming, our method includes a machine learning-based data augmentation module.

This module, based on a *deep neural network* (DNN), can efficiently approximate the underlying molecular system and turns out to be key for an effective bound optimization. Machine learning has already found some applications in MC. For example, in [12] it was demonstrated the possibility of training deep learning-based detectors that perform well without any prior knowledge of the underlying channel model — as often happens in MC. In [13], the authors introduced a neural network-based technique for modeling the molecular multiple-input multiple-output channel, and recently, in [14], different supervised machine learning methods were proposed to estimate the distance between transmitter and receiver in an MC system.

We illustrate our proposed data-driven approach on a system composed of two genetically engineered cells [15], for which in-silico simulations are produced.

The paper is organized as follows. In Sec. II we summarize

978-1-7281-8298-8/20/$31.00 ©2020 IEEE

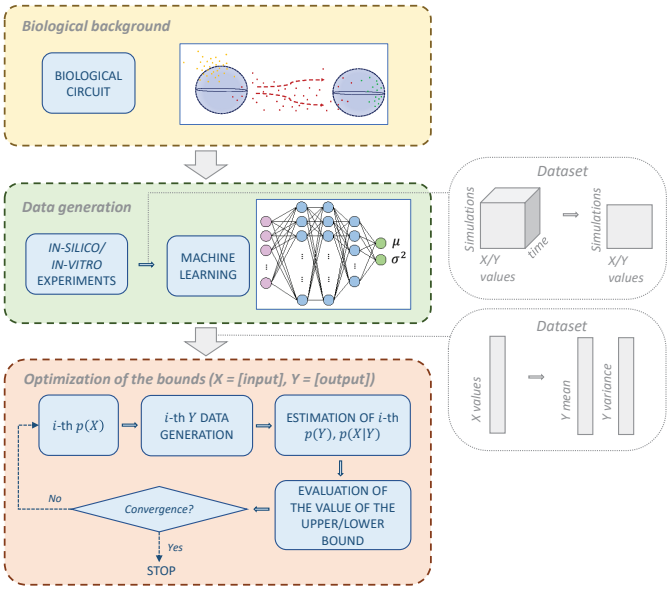


Fig. 1: Overview of the proposed methodology, where X is the source message, Y the destination message

the proposed methodology to optimize the bounds on the channel capacity. In Sec. III we describe the biological model representing the system under investigation and the proposed machine learning-based data augmentation module. In Sec. IV we summarize the information theory background of the proposed algorithm. The optimization of the channel capacity bounds is the subject of Sec. V, where the proposed algorithm is detailed. Last, in Sec. VI we discuss the results and then in Sec. VII we conclude the paper.

II. Methodological overview

The aim of this section is to provide a general overview of the developed methodology. Figure 1 visually resumes the main steps of this work.

Given a biological circuit, the proposed methodology has the objective of optimizing its bounds on the channel capacity. To reach this goal, an *a priori* dataset of source messages and corresponding received signals for the target biological system is required. Such a dataset can be obtained through in-silico or in-vitro experiments.

For this paper, we consider the running example of a biological circuit defined by two prokaryotic cells [15], and we perform in-silico simulations to generate a dataset containing time-dependent values of the destination messages (outputs) given different source messages (inputs).

Given the fact that biological simulations are heavily time consuming, we cannot rely on on-demand simulations during the execution of the proposed bounds estimation iterative algorithm. Instead, we use the set of previously generated simulations as input of a *data augmentation* module [16] to efficiently generate new samples, fitting any required distribution without the need to run any new simulation. Through the combination of machine learning and statistical sampling, the data augmentation module allows for the high-performance generation of realistic outputs. The proposed generation technique also models the inherent stochasticity of the biological process, totally replacing in-silico simulations for the target interval range. The biological circuit used for this work and the data augmentation procedure are detailed in Sec. III.

In addition to the dataset of input/output samples, the iterative algorithm takes as input a parametric family of distributions $\mathcal{F}$. At convergence, the algorithm determines the optimal bounds on the channel capacity and the corresponding input distributions $f_{lower}, f_{upper} \in \mathcal{F}$. Both the formulas for the upper and the lower bound rely on the distributions of the source and the destination messages, which are in turn estimated from the data, using the data augmentation module. Details on the theoretical formulation of the upper and lower bounds are given in Sec. IV, while the iterative algorithm is described in Sec. V.

We remark that the proposed methodology is independent with respect to the considered biological circuit and its modeling. The proposed algorithm only requires a previously available dataset of realistic input/output samples and does not rely on new simulations/experiments during its execution.

III. Data generation

In this section we briefly introduce the biological system that has been simulated for the generation of in-silico data, which corresponds to our running example for the following methodology description and results. We then detail the machine learning-based data augmentation of the in-silico simulations, which corresponds to the first step of the proposed methodology.

A. The in-silico biological system

Figure 2 shows a scheme of the biological system described in [15], used as a running example throughout the paper. The system is composed of two prokaryotic cells of the same size collocated at a distance of 30 μm in an environment of infinite volume. Both cells have a unique function. That of the first cell is to transmit a message through the biological communication system, while that of the second cell is to act as receiver. The *source message* is a concentration of the protein inducer Isopropyl β-d-1-thiogalactopyranoside (*IPTG*). Different levels of IPTG concentration represent different symbols in the modulation of the source message. The instantaneous injection of a concentration value of IPTG in the transmitter cell induces a series of chemical reactions in the cell itself that, in turn, releases in the extracellular space signaling molecules of acyl-homoserine lactone (*AHL*). The AHL diffuses with Brownian motion in the space surrounding the two cells [17]. The receiver cell membrane receptors capture the diffused AHL molecules. Chemical reactions induced in the receiver cell due to the presence of AHL lead to the production of a level of concentration of the green fluorescent protein *GFP*, the *destination message*. Details about the chemical reactions occurring in the two cells can be found in [15]. In-silico simulations of this MC system were performed by using the

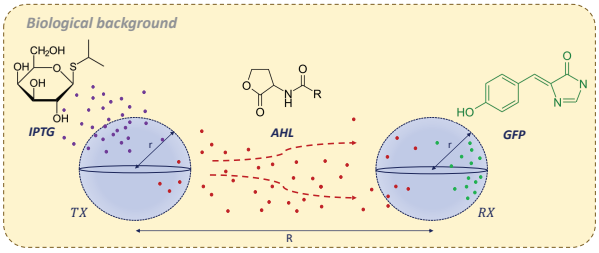


Fig. 2: Scheme of the considered biological circuit

MATLAB Simbiology Toolbox. The range of input IPTG concentrations is varied from $1.5 \cdot 10^6$ nM to $1.5 \cdot 10^8$ nM with a step size of $1.5 \cdot 10^6$ nM.

In these in-silico experiments, we simulate a Concentration Shift Keying (CSK) modulation [18] and assume a memoryless channel, i.e. the absence of inter-symbol interference. These hypothesis allow us to simplify simulations and evaluate the bounds in the ideal condition of independent samples for the received signal.

We remark that the proposed methodology is valid also for datasets derived by more complex scenarios, if appropriately pre-filtered to remove the effects of the memory of the channel and mitigate the potential impact of inter-symbol interference on the data.

The simulations are stochastic to capture the effect of the noise given both by the diffusion process and by the probabilistic occurrence of the chemical reactions. On the other hand, the stochasticity would add uncertainty to the simulations and, consequently, to our calculations and results. Thus, we repeat all the simulations $N = 100$ times to empirically characterize the effect of the noise on the data. To the best of our knowledge, an analytical statistical characterization of the occurrence of the chemical reactions has not been introduced in the literature yet. Our data-driven method allows an estimation of the channel capacity without the need of an analytical statistical modeling.

Our output dataset can be seen as a 3D tensor, where the three dimensions are: the time instant when the i-th concentration is recorded, the concentration value in that instant, and the different simulations made for each input concentration. To obtain more tractable data, we reduce the dataset from three to two dimensions, by considering for each input concentration value only the instant of time corresponding to the maximum value of the output concentration. The input dataset has been pre-processed in the same way, obtaining a 2D tensor over maximum IPTG concentrations and simulations.

In the following, *X* stands for the input and *Y* for the output, i.e., IPTG and GFP in the considered biological system. Furthermore, we consider X and Y as discrete random variables, as in our running example the input, and by consequence the output, concentration values are discrete.

B. A machine learning approach to augment the dataset

Iteratively optimizing the channel capacity requires evaluating the output distribution for any possible input distribution in the target family $\mathcal{F}$. Moreover, having data corresponding to only a limited set of input/output pairs (e.g., 100 different IPTG concentration levels in our running example) could lead to an imprecise estimation of the distributions (and, consequently, of the bounds on the channel capacity). In-silico simulations are time consuming, thus cannot be run at each iteration of the algorithm and with a high density in the interval of interest. To address both these problems and ensure the generality of our method, we propose to utilize a machine learning-based data augmentation approach.

Once trained, a machine learning model output is intrinsically deterministic. As we generate data that are affected by stochastic noise, we want our machine learning model to reflect this stochasticity. To deal with this, we frame our problem as a *probability density estimation* problem [19], as detailed next.

First, we use multiple in-silico simulations for each input $x \in X$ to statistically characterize the corresponding outputs. In the running example, 100 simulations have been executed for each x. Then, we assume a particular shape for the output distribution $p(y \mid x)$ and we train the model to predict its parameters. In our case, given an IPTG concentration, the corresponding GFP value can be approximated by a Gaussian distribution[1]. Given an input value x, the model is therefore trained to predict *both* the mean μ_{GFP} and the variance σ^2_{GFP} of the corresponding stochastic output. Finally, a predicted distribution is used to sample $y \sim \mathcal{N}\left(\mu_{GFP}, \sigma^2_{GFP}\right)$, thus allowing an arbitrary number of stochastic outputs to be sampled for each input. This process is efficient and takes only a fraction of the time required to gather new simulations.

The proposed methodology does not depend on a specific machine learning model. In this work, we use a *deep neural network*-based model [20], which reported the best performance in a preliminary analysis[2]. We can speculate that a DNN can effectively model the large number of nonlinearities and interdependencies characterizing the biological circuit, becoming an efficient proxy for the underlying biological model once trained.

The network used in this work is a feedforward DNN with L (hidden) layers and a number of neurons in each layer $N_i = \frac{N_1}{2^{i-1}}$, with N_1 being the number of neurons in the first layer and $i = 1, \ldots, L$ the different layers. With this design, only two hyper-parameters define the network structure and need to be optimized. The network is trained to minimize the mean squared error (MSE), which is equivalent to using maximum likelihood estimation assuming that the data is from a normal distribution [20]. For training we use gradient-based optimization with RMSprop ($\epsilon = 0.001$), a first-order stochastic optimizer, and early stopping (for regularization) up to 1000 epochs. We use a 80/10/10 data split for training/validation/test sets. We use k-fold cross-validation to deal with data scarcity and increase the robustness of the design

[1]This has been validated through in-silico experimental results.

[2]The choice of the model has been made based on a preliminary comparison of different machine learning algorithms, including support vector machines (SVM), random forest (RF), and feedforward deep neural network.

choices on the validation set. Hyper-parameter optimization on the validation set led to the final design of our network in terms of number of layer/units. The network has been implemented with Tensorflow/Keras [21].

The final trained network reports a mean absolute error of $1.4 \cdot 10^4$ with respect to the true test values of μ_{GFP}. To get a sense of the network accuracy, we should take into account the inherent variability of the simulation outputs for a given input, characterized by a mean variance of $\approx 2 \cdot 10^{10}$. We can verify that the predicted value always falls in the interval defined by the first and third quartile of the output distribution of the simulations for a certain input. Therefore, predictions appear to be accurate when compared to the underlying stochastic in-silico model.

We use the trained network to predict 10^4 equally spaced output concentrations (means and variances) given the same number of input concentrations in the interval of interest, thus significantly increasing the density of our predictions with respect to the starting ones (10^2). These predictions are used to efficiently sample any required input distribution during the optimization of the bounds on the capacity, as explained next.

IV. Upper and Lower Bounds on the Channel Capacity

In this section we introduce the information theory background for the definition of the bounds, starting from the probability mass functions (pmf) of the input and output concentrations $P(X)$ and $P(Y)$ and the conditional pmfs.

A. The lower bound

Given a discrete random variable X its entropy is defined as [22]:

$$H(X) = -\sum_{j=0}^{M-1} P(x_j) \log_2 P(x_j), \tag{1}$$

where M is the sample space of X. The conditioned entropy of the input X given the output Y is:

$$H(X \mid Y) = -\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} P(y_i) P(x_j \mid y_i) \log_2 P(x_j \mid y_i), \tag{2}$$

where N is the sample space of Y.

The MI of two random variables is a measure of the mutual dependence between them. For this reason, the definition of MI is [22]:

$$I(X,Y) = H(X) - H(X \mid Y). \tag{3}$$

Consequently, the channel capacity is defined as:

$$C = \max_{P(X) \in \mathcal{P}(\mathcal{X})} I(X,Y). \tag{4}$$

As defined in equation (4), channel capacity can be estimated correctly only if all the possible pmfs are explored at the input, which is clearly not feasible in practice.

However, Eq. (4) leads to a natural lower bound on the capacity, since any input distribution $P(X)$ leads to the inequality:

$$C \geq I(X,Y). \tag{5}$$

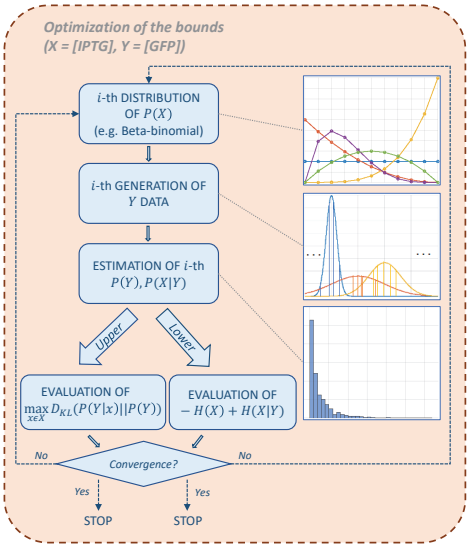


Fig. 3: Detailed overview of the data-driven methodology implemented to optimize the bounds on the channel capacity

B. The upper bound

The definition of Kullback-Leibler divergence (D_{KL}) leads to a dual formula for the MI, that is [23]:

$$I(X,Y) = D_{KL}(P(X,Y) \parallel P(X)P(Y)). \tag{6}$$

Furthermore, it is possible to write:

$$D_{KL}(P(Y \mid x) \parallel P(Y)) = \sum_{y \in Y} P(y \mid x) \log_2 \frac{P(y \mid x)}{P(y)}. \tag{7}$$

From Eq. (6) and Eq. (7) we can derive a dual formula to express the capacity for amplitude-constrained channels [23]:

$$C = \min_{P(Y) \in \mathcal{P}(\mathcal{Y})} \max_{x \in X} D_{KL}(P(Y \mid x) \parallel P(Y)). \tag{8}$$

Therefore, every choice of a distribution $P(Y)$ leads to an upper bound on the channel capacity [24]:

$$C \leq \max_{x \in X} D_{KL}(P(Y \mid x) \parallel P(Y)). \tag{9}$$

V. Optimization of the Bounds

In the following we describe the proposed iterative algorithm that, given as input a dataset $\mathcal{S} = (X,Y)$ and a parametric distribution, finds the parameters of the input distribution optimizing the bounds.

The algorithm is independent with respect to the choice of the family of distributions to be explored. Such a choice should be driven by the specific features of the considered system. For the experiments in this work, we use the *Beta-binomial distribution*. The rationale behind this choice is detailed next.

A. The Beta-binomial distribution

The choice of the beta-binomial distribution $X \sim Beta\text{-}binomial(a,b)$, $a > 0$, $b > 0$, for this work has multiple motivations. First, it has been demonstrated that the input probability density function (pdf) that maximizes the

channel capacity for a Binomial channel is discrete [25]. Moreover, we need to choose a distribution defined on a limited interval coinciding with the minimum/maximum input concentrations (i.e., $1.5 \cdot 10^6$ nM – $1.5 \cdot 10^8$ nM in our setting). Furthermore, as it would be impossible to try as input all the existing $P(X)$, we need to select a parametric distribution characterized by high heterogeneity in terms of shapes with respect to the parameters. This allows testing different trends and covering as many cases as possible. As the pdfs of the Beta distribution family, the Beta-binomial distributed pmfs can be easily shifted in any desired positive interval [26].

B. The optimization algorithm

To account for the *implicit* characterization of the MI function with respect to the input distribution parameters, which stems from a data-driven approach, the proposed algorithm requires a *derivative-free* optimization method.

Among existing derivative-free optimization methods, the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [27] is an evolutionary algorithm that minimizes a function based on the iterative update of the covariance matrix of the multivariate Gaussian distribution from which candidate solutions are chosen. It has been shown [27], [28] that the CMA-ES outperforms other derivative-free optimization algorithms in terms of total number of function evaluations needed to reach the optimal value. Furthermore, it exhibits notable invariant properties, in particular to order preserving transformations of the objective function value. At each generation g_i of the CMA-ES, a set of possible solutions is evaluated, and, based on the corresponding values of the objective function, the values that lead to the best fitness are used to update the mean and the covariance matrix of the successive generation g_{i+1}, while the others are discarded.

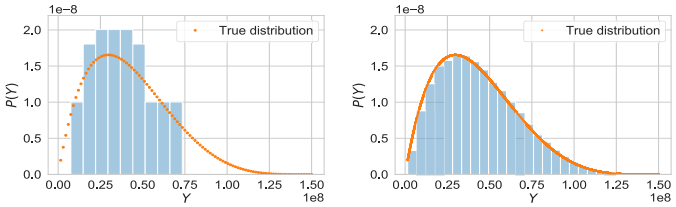
The algorithm is run two times to optimize, respectively, the upper bound (Eq. (9)) and the lower bound on the channel capacity (Eq. (5)). As discussed in Sec. IV, both equations depend (directly or indirectly) on $P(X)$. Therefore, we optimize the bounds as function of the distribution parameters (a and b for the considered case). In the end, we obtain optimal bounds and the corresponding input pmfs $P(X)$.

C. Bounds optimization

Figure 3 provides an overview of the proposed methodology to optimize the bounds on the channel capacity of a biological circuit, given a dataset generated as detailed in Section III.

At each iteration (generation g_i) of the optimization algorithm, we calculate the MI for the pmfs $P(X)$ of the input message (i.e., in the running example, different concentration levels of IPTG) which belong to that generation. This is done through a *data-driven* approach.

First, we generate a pair of input/output data distributions matching each considered pmf through a sampling procedure starting from the augmented dataset, as explained in Section III-B. Data distributions are generated for each pmf ensuring that the total number of samples is constant ($n = 10^7$ in our experiments) and well-distributed over the target concentration ranges. Specifically, given p_i the probability corresponding to the i-th concentration value, $k_i = n \cdot p_i$ samples are generated for the i-th concentration value. The total number of samples per iteration has been chosen as a trade-off between a pmf approximation accuracy and computational efficiency.



(a) (b)

Fig. 4: Example of distribution estimation *without* data augmentation (a) and *with* data augmentation (b)
Results are based on data described in Section III

The algorithm is initialized with a uniform $P(X)$, i.e., $a = b = 1$. During each iteration, both $P(Y)$ and $P(X \mid Y)$ are estimated from the dataset using *histograms*. The number and the width of the bins are equal for each distribution at the same iteration of the algorithm. The width of each bin does not thereby affect the calculation of the bounds. The total number of bins for each pmf is computed applying the Doane's formula [29] to the samples of X and Y and by choosing the lowest number between the two at the i-th iteration. This is motivated by the fact that the Doane's formula defines an upper bound on the required number of bins to reach a desired precision. Hence, we obtain an estimation of the pmfs with an acceptable degree of precision without adding complexity to the calculations.

Figure 4 shows the combined effect of the data augmentation and the histogram-based estimation of $P(Y)$. Without data augmentation (Fig 4a), the number of bins is low, and the resulting approximation deviates significantly from the true distribution. Data augmentation leads to a larger number of bins and a more accurate estimation (Fig 4b).

The evaluation of $P(Y)$ is fundamental to estimate $P(X \mid Y)$. In fact, for each bin of $P(Y)$, a histogram $P(x \mid y)$ is built by selecting all the x values that correspond to the y values included in the considered bin of $P(Y)$. $P(Y \mid X)$ in equation (9) is obtained via the Bayes theorem [22].

The same iterative algorithm used to minimize the upper bound is also used to maximize the lower bound. To do this, we minimize the opposite of Eq. (3).

VI. Results and Discussion

Results in the following are based on the scenario described in Section III. Figure 5 exhibits the progressive convergence on the minimum of the upper bound on the channel capacity as the number of generations of the CMA-ES algorithm increases. The a and b parameters of the input Beta-binomial distribution are adjusted at each iteration to select a population of points

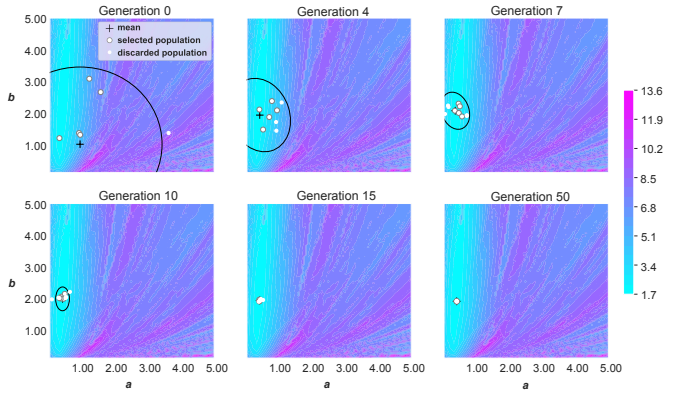
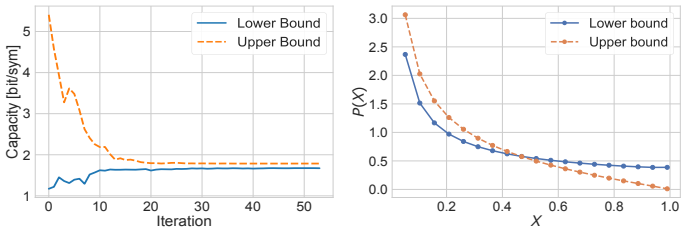


Fig. 5: Progressive convergence of the CMA-ES algorithm



(a) (b)

Fig. 6: (a): Optimized bounds on the channel capacity
(b): Standardized input pmf corresponding to the final optimized bounds. The number of represented points is less than 10^4 (i.e. the number of input IPTG concentrations) to have a clearer picture

that is gradually less scattered and closer to the minimum of the function. The algorithm does not ensure that the reached solution is the global minimum, but it has been shown that it can approximate one of the best local minima of the considered function [30]. In Fig. 5, the mean represents the best solution at each generation, while the ellipse encloses 95% of the sampled points. The larger circled dots represent the sampled solution with the best fitness at each generation. They are used to update the mean and covariance matrix for the next generation. The smaller non-circled dots are the worst sample solutions, which are discarded at each generation. Note that, without the dataset, it would not have been possible to evaluate the function to be minimized. In fact, we do not have an analytical formulation of the bounds that depends only on the a and b the parameters of $P(X)$. In Fig. 5 the function has been evaluated numerically.

Figure 6 shows progressive optimization of the bounds and the input pmfs corresponding to the best parameters of the upper and of the lower bound. In particular, Fig. 6a shows how the bounds become closer as the optimization progresses. The main steps towards convergence appear to be made in the first few steps, reaching convergence around the 50^{th} generation. Results show that the achievable channel capacity of the considered biological system lies within the range $[1.67, 1.71]$ [bit/symbol].

As shown in Fig. 6b, the $P(X)$ corresponding to the optimized bounds are very similar, due to the fact that the resulting bounds are very close. Both pmfs converge to a shape that recalls the one of an exponential distribution. This result suggests that the $P(X)$ corresponding to the actual value of the channel capacity decreases non-linearly as the IPTG concentration value increases. Thus, the optimal pmfs for our running example privilege lower values of IPTG concentration. This may be due to the saturation of the measured average value of the GFP concentration level in the receiver cell as we increase the IPTG concentration value in the transmitter cell. The resulting optimized values are $a = 0.47$ and $b = 1.90$ for the upper bound, $a = 0.35$ and $b = 0.98$ for the lower bound.

VII. Conclusion

In this work, we described a data-driven methodology to optimize the channel capacity bounds of an MC system. Our approach only requires a dataset of output messages conditioned to inputs, which can be generated via in-vitro/in-silico experiments, and a family of distributions as domain of the optimization procedure. The starting dataset is augmented via machine learning to increase the density of the output measurements in the interval of interest. The augmented dataset constitutes the basis for optimizing the capacity bounds through an iterative approach exploiting a derivative-free optimization algorithm (CMA-ES) and efficient sampling of arbitrary distributions. As output, the algorithm finds a relatively small interval in which the value of the capacity falls for the considered molecular system and the shape of the relative pmf of the input data.

Future work should investigate the application of the proposed methodology to a more realistic biological scenario. The effects on the information exchange performance of a channel with memory and of the limited amount of resources present in a cell should be considered. Moreover, a more in-depth analytical investigation of the obtained results should be carried out.

References

[1] I. F. Akyildiz, F. Brunetti, and C. Blázquez, "Nanonetworks: A new communication paradigm," *Computer Networks*, vol. 52, no. 12, pp. 2260–2279, 2008.

[2] O. B. Akan, H. Ramezani, T. Khan, N. A. Abbasi, and M. Kuscu, "Fundamentals of molecular information and communication science," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 306–318, 2017.

[3] G. D. Forney and G. Ungerboeck, "Modulation and coding for linear gaussian channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2384–2415, 1998.

[4] V. Jamali, A. Ahmadzadeh, W. Wicke, A. Noel, and R. Schober, "Channel modeling for diffusive molecular communication—a tutorial review," *Proceedings of the IEEE*, vol. 107, no. 7, pp. 1256–1301, 2019.

[5] N. Farsad, H. B. Yilmaz, A. Eckford, C.-B. Chae, and W. Guo, "A comprehensive survey of recent advancements in molecular communication," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1887–1919, 2016.

[6] Q. Liu and K. Yang, "Channel capacity analysis of a diffusion-based molecular communication system with ligand receptors," *International Journal of Communication Systems*, vol. 28, no. 8, pp. 1508–1520, 2015.

[7] M. Pierobon and I. F. Akyildiz, "Capacity of a diffusion-based molecular communication system with channel memory and molecular noise," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 942–954, 2013.

[8] H. Awan and C. T. Chou, "Improving the capacity of molecular communication using enzymatic reaction cycles," *IEEE transactions on nanobioscience*, vol. 16, no. 8, pp. 744–754, 2017.

[9] Z. Cheng, Y. Zhu, K. Chi, Y. Li, and M. Xia, "Capacity analysis for diffusive molecular communication with isi channel," *Nano communication networks*, vol. 13, pp. 43–50, 2017.

[10] P. Sadeghi, P. O. Vontobel, and R. Shams, "Optimization of information rate upper and lower bounds for channels with memory," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 663–688, 2009.

[11] S. Verdú, "Empirical estimation of information measures: A literature guide," *Entropy*, vol. 21, no. 8, p. 720, 2019.

[12] N. Farsad and A. Goldsmith, "Detection algorithms for communication systems using deep learning," *arXiv preprint arXiv:1705.08044*, 2017.

[13] C. Lee, H. B. Yilmaz, C.-B. Chae, N. Farsad, and A. Goldsmith, "Machine learning based channel modeling for molecular mimo communications," in *2017 IEEE 18th international workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2017, pp. 1–5.

[14] F. Gulec and B. Atakan, "Distance estimation methods for a practical macroscale molecular communication system," *Nano Communication Networks*, p. 100300, 2020.

[15] C. Harper, M. Pierobon, and M. Magarini, "Estimating information exchange performance of engineered cell-to-cell molecular communications: a computational approach," in *INFOCOM*, 2018, pp. 729–737.

[16] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *J. Comput. Graph. Stat.*, vol. 10, no. 1, pp. 1–50, 2001.

[17] I. Karatzas and S. E. Shreve, "Brownian motion," in *Brownian Motion and Stochastic Calculus*. Springer, 1998, pp. 47–127.

[18] M. S. Kuran, H. B. Yilmaz, T. Tugcu, and I. F. Akyildiz, "Modulation techniques for communication via diffusion in nanonetworks," in *ICC*, 2011, pp. 1–5.

[19] J. Walters-Williams and Y. Li, "Estimation of mutual information: A survey," in *International Conference on Rough Sets and Knowledge Technology*. Springer, 2009, pp. 389–396.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[22] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[23] I. Csiszar and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[24] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2426–2467, 2003.

[25] R. D. Wesel, E. E. Wesel, L. Vandenberghe, C. Komninakis, and M. Medard, "Efficient binomial channel capacity computation with an application to molecular communication," in *ITA Workshop*. IEEE, 2018, pp. 1–5.

[26] A. K. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications*. CRC press, 2004.

[27] N. Hansen, "The cma evolution strategy: a comparing review," in *Towards a new evolutionary computation*. Springer, 2006, pp. 75–102.

[28] N. Hansen, A. Auger, R. Ros, S. Finck, and P. Pošík, "Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009," in *Conference Companion on Genetic and Evolutionary Computation*, 2010, pp. 1689–1696.

[29] D. P. Doane, "Aesthetic frequency classifications," *The American Statistician*, vol. 30, no. 4, pp. 181–183, 1976.

[30] N. Hansen and A. Ostermeier, "Convergence properties of evolution strategies with the derandomized covariance matrix adaptation: The (/i,)-es," *Eufit*, vol. 97, pp. 650–654, 1997.